ChatGPT

Qwen4-Max-Preview 徹底調査レポート

エグゼクティブサマリ

2026年11月15日時点で、**Qwen4-Max-Preview は実在する公式モデル**です。Alibaba Cloud Model Studio / QwenCloud の公式ドキュメントに `qwen4-max-preview` として掲載されており、別名として `qwen4.0-max-preview` も利用できます。Alibaba はこれを「Qwen ファミリーの次世代フラッグシップ・マルチモーダルモデル」と位置付けています。公開開始は **2026年11月11日 10:00（UTC+8）**、Preview は **2026年11月25日 24:00（UTC+8）まで**で、Alibaba はこの Preview モデルを期間終了後に retire（提供終了）すると明記しています。したがって、これは通常の「安定版 API リリース」ではなく、**期間限定の評価用 Preview** とみなすのが正確です。 [1]

最も重要な注意点は、性能そのものよりも**提供条件**です。Qwen4-Max-Preview は通常の従量課金 API では提供されず、**Token Plan 限定**です。国際向け Alibaba Cloud では Token Plan は現時点で **Singapore リージョンのみ**、中国向けでは Beijing リージョンで提供されています。また Token Plan の利用規約は、Claude Code、Cursor、Qwen Code、Qoder、OpenClaw 等の**対話的な coding/agent tool 内での利用だけを許可**しており、自動バッチ、独自スクリプト、SaaS/アプリのバックエンドからの API 呼び出しは禁止されています。技術的には OpenAI / Anthropic 互換 API を持ちますが、「API がある＝自由にプロダクトへ組み込める」わけではありません。 [2]

仕様面は非常に大きく、**1M token context、最大 393,216 token の reasoning＋最終出力共通予算、text/image/video input、text output、Function Calling、Web Search、built-in tools、JSON Object / JSON Schema** を備えます。Thinking はデフォルト ON で `reasoning_effort=low|medium|xhigh`、既定値は `xhigh` です。これは Qwen3.8-Max の最大出力 128K を大幅に上回ります。 [3]

一方、**Alibaba は Qwen4-Max-Preview の正式なベンチマーク表、パラメータ数、技術論文、モデルカード、ダウンロード可能な重みを現時点で公表していません**。したがって「GPT-6 Astra より強い」「Qwen3.8-Max を超えた」といった主張を公式データから立証することはできません。独立系 Stealth Models の初期測定では StealthMark が **58.8/100** で、同じ評価系では Qwen3.8-Max-0902 の **62.4**、Gemini 4 Argon の **67.4**、GPT-6 Astra の **74.6** を下回っています。ただし StealthMark は独自ベンチマークであり、MMLU、GPQA、SWE-bench 等の標準評価とは同列に扱うべきではありません。 [4]

採用判断としては、現時点では「評価・探索には非常に興味深いが、本番依存先には不適切」です。理由は、Preview が2026年11月25日に終了すること、安定版への移行保証がないこと、Token Plan がアプリバックエンドを禁止していること、公式ベンチマークと明示的 SLA/長期価格が不足していることです。業務用 coding agent や研究評価には試す価値がありますが、プロダクト組込みなら安定版 Qwen3.8-Max、DeepSeek V4.1、Gemini 4 Argon、GPT-6 系などを比較対象に置くべきです。 [5]

実在性、正式名称、Qwen 系統との関係

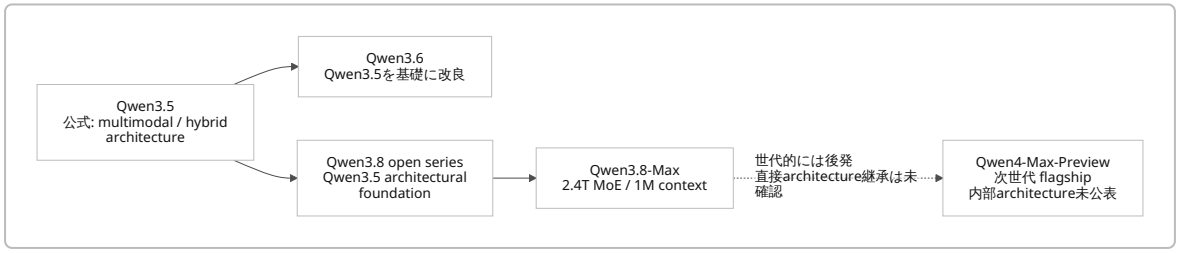
確認済みの正式情報

項目	2026-11-15時点の確認結果	判定
正式表示名	Qwen4-Max-Preview	**確認済み**

項目	2026-11-15時点の確認結果	判定
API model ID	`qwen4-max-preview`	**確認済み**
Alias	`qwen4.0-max-preview`	**確認済み**
提供元	Alibaba Cloud Model Studio / Qwen	**確認済み**
公開日	2026-11-11	**確認済み**
Preview開始	2026-11-11 10:00 UTC+8	**確認済み**
Preview終了	2026-11-25 24:00 UTC+8	**確認済み**
Preview終了後	Preview model を retire	**確認済み**
正式版 Qwen4-Max の発売日	未公表	**未確認**
パラメータ数	未公表	**未確認**
Architecture	未公表	**未確認**
Open weights	公開確認できず	**未確認／現状なし**
技術論文/model card	公開確認できず	**未確認／現状なし**

上記の名称・期間・alias は Alibaba の公式モデルページと Token Plan ページで直接確認できます。特に「Preview 後 retire」と明示されているため、一般的な「Preview＝同じ ID のまま安定版へ昇格」という意味ではありません。 6

「Qwen4」という世代名は公式に使われていますが、Qwen4-Max-Preview の**内部アーキテクチャが Qwen3.8-Max の直接後継なのか、Qwen3.8 の 2.4T MoE と同系列なのかは公式資料では確認できません。**対照的に、Qwen3.8 は公式 GitHub で Qwen3.5 の architectural foundation を基礎とすると説明され、Qwen3.8-Max は 2.4 trillion parameter MoE と明記されています。Qwen4 に同じ数字を流用してはいけません。 7

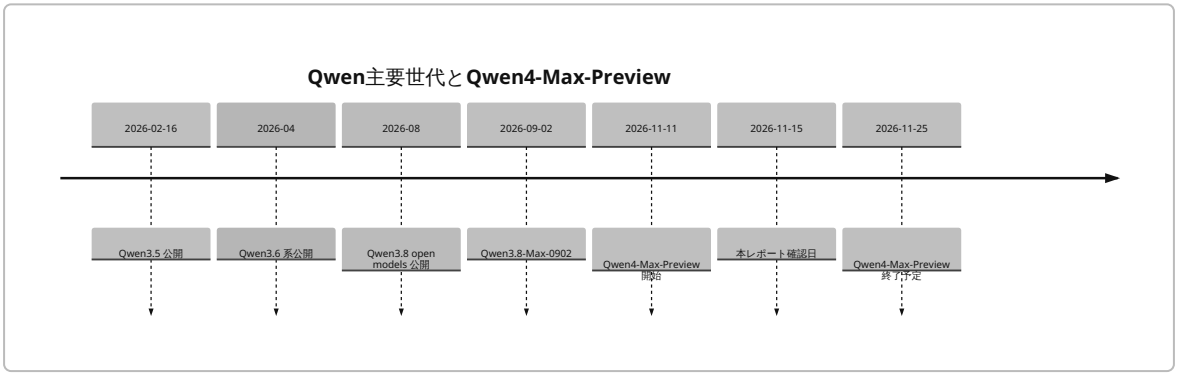


Qwen3.8-Max と Qwen4-Max-Preview を混同しやすい最大の原因は、どちらも「Max」「1M context」「reasoning」「native visual understanding」「tool ecosystem」を持つことです。しかし Qwen3.8-Max は安定提供され、従量課金 API も存在し、2.4T MoE という技術情報と価格が公開されています。一方 Qwen4-Max-Preview は Token Plan 限定で、内部規模も通常 API 価格も非公開です。 8

比較項目	Qwen4-Max-Preview	Qwen3.8-Max
状態	期間限定 Preview	Stable
Context	1M	1M
Max output	393,216（reasoning込み）	128K

比較項目	Qwen4-Max-Preview	Qwen3.8-Max
Text input	○	○
Image input	○	○
Video input	○	○
Thinking	○	○
Function Calling	○	○
Built-in tools	○	○
公開パラメータ数	**未公表**	2.4T MoE
通常従量 API	×	○
Open weights	確認できず	Max は hosted proprietary
本番 backend	Token Plan 規約上不可	通常APIで可

Qwen3.8-Max の公式 GitHub は API 価格を input $2 / output $6 / cached $0.25 per 1M tokens としています。Qwen4-Max-Preview に同等の per-token price は公開されていません。[9]



Qwen3.5〜3.8 の公開日は Qwen 公式 GitHub の release history、Qwen4-Max-Preview は Alibaba Cloud の release log に基づきます。[10]

提供地域、アクセス方法、API、料金

地域と利用資格

国際版 Alibaba Cloud の Token Plan Personal / Team は、2026年11月15日時点で **Singapore (`ap-southeast-1`) のみ**です。Alibaba Cloud の一般的な Model Studio 自体は Tokyo、Frankfurt、Virginia 等も持ちますが、それは Qwen4-Max-Preview の Token Plan 提供地域を意味しません。したがって、日本在住ユーザーが International Site を使う場合も Qwen4-Max-Preview のエンドポイントは現状 Singapore です。[11]

中国サイト側では Token Plan が China (Beijing) で提供されています。中国版と International 版では Base URL も課金通貨も異なるため、ドキュメントを混在させないことが重要です。[12]

API endpoint

International / Singapore の公式 endpoint は以下です。[13]

```
OpenAI-compatible:
https://token-plan.ap-southeast-1.maas.aliyuncs.com/compatible-mode/v1

Anthropic-compatible:
https://token-plan.ap-southeast-1.maas.aliyuncs.com/apps/anthropic
```

model は以下です。

```
qwen4-max-preview
```

Token Plan API Key は通常の Model Studio API key `sk-...` ではなく、**専用の `sk-sp-...` key** を使います。両者は相互利用不可です。 13

OpenAI Chat Completions、OpenAI Responses、Anthropic Messages の互換性が公式に明記されています。SDK 固有の専用 Qwen SDK が必要なわけではなく、OpenAI / Anthropic 互換 client を Base URL 差し替えで利用できます。 14

サインアップの流れ

International site では、Alibaba Cloud account を作成し、Token Plan Personal または Team を購入、Token Plan 専用 API key を発行し、Singapore の Base URL と `qwen4-max-preview` を coding/agent tool に設定します。Team の場合は seat を member に割り当てた後に individual key が生成されます。 15

ただし重要なのは、**この API key を Python script や独自 backend から直接叩く通常の「API integration」は Token Plan 利用規約上の許可用途ではない**という点です。Alibaba は自動スクリプト、custom application backend、non-interactive batch を明示的に禁止しています。 16

料金

International Personal Edition は現在以下です。 17

Plan	通常月額	2026-11-30までの新規限定価格	月間 Credits
Lite	$6	$3	20,000
Essential	$10	$5	45,000
Standard	$18	—	80,000
Pro	$43	—	200,000

追加 bundle は **$15 / 20,000 Credits / month**、最大5個です。Token Plan 自体には free trial Tokens / free Credits はありません。 18

Team は Standard $20/seat/month 25,000 Credits、Advanced $75/100,000 Credits、Premium $180/250,000 Credits です。 19

Qwen4-Max-Preview は Preview 中、通常 Credits consumption の **20%**、さらに 22:00–08:00 UTC+8 はそこから 60% 引き、**通常の8%相当**とされています。 20

ただし Credits は token そのものではありません。入力 token、cached token、output token、thinking、tool call 等により request ごとの消費が動的に決まります。したがって「1 Credit = X tokens」という固定換算はできず、**Qwen4-Max-Preview の per-million-token 公開価格もありません。** [21]

技術仕様と実用上の特性

公式に確認できる主要仕様は次の通りです。 [3]

仕様	Qwen4-Max-Preview
Context window	**1,000,000 tokens**
Maximum completion budget	**393,216 tokens**
Reasoning + answer	同一 completion budget を共有
Text input	○
Image input	○
Video input	○
Audio input	**公式対応確認なし**
Text output	○
Native image generation	×／別モデル
Native audio output	×／別モデル
Function Calling	○
Built-in web search	○
Built-in tools	○
JSON Object	○
JSON Schema	○
OpenAI Chat Completions	○
OpenAI Responses	○
Anthropic Messages	○
Thinking default	ON
`reasoning_effort`	`low`, `medium`, `xhigh`
Default effort	`xhigh`
Max image pixels	16M pixels/image
Max images	2,048
Max video duration	2時間
Max video size	2 GB

仕様	Qwen4-Max-Preview
Max videos	64

「code」は独立した input modality ではなく text task の一種です。また Qwen4-Max-Preview 自体が音声モデルという記述はありません。Alibaba は speech recognition / TTS / realtime audio を Qwen Audio 系の別モデルとして提供しています。 17

API の注意点として、Qwen4-Max-Preview から OpenAI-compatible `max_tokens` は `max_completion_tokens` と同義になり、**reasoning token と最終 answer の合計**を制限します。たとえば `max_tokens=8192` とすると、長い thinking が大半を使い切り、回答本体が短くなる可能性があります。 22

Chat Completions の multi-turn では `preserve_thinking` がデフォルト有効で、過去の thinking は `reasoning_content` として維持する設計です。 14

レイテンシと推論環境

公式の tokens/sec、TTFT、GPU 型式、GPU 数、VRAM 要件は公表されていません。 また model weights / self-host instructions も確認できないため、オンプレミス GPU 導入に必要なハードウェアを算定する根拠はありません。現在の公式な利用形態は hosted Token Plan と考えるべきです。 14

第三者 Stealth Models は Qwen4-Max-Preview について約 **66 tokens/s** を測定していますが、これは Alibaba の保証値ではなく、接続先、負荷、prompt、reasoning effort により変化する観測値です。信頼度は「中〜低」と評価します。 23

実用上の強みは、巨大 context、非常に大きい completion budget、画像・長時間動画、tool calling、web、structured output を一モデルで扱える点です。Alibaba 自身は coding、office productivity、frontend design、complex agentic tasks を主要用途として挙げています。 14

弱みは、Preview 固有の寿命、`xhigh` default による思考量・Credits 消費、公式性能データの不足、backend 禁止、そして初期ユーザーから報告されている tool-call 周辺の挙動不安定性です。 24

性能、ベンチマーク、競合比較、利用者評価

公式ベンチマークの状況

2026年11月15日までに確認した Alibaba/Qwen の Qwen4-Max-Preview 公式ページには、MMLU、GPQA、AIME、SWE-bench、Terminal-Bench 等の**具体的スコア表は掲載されていません**。パラメータ数や学習 corpus、論文へのリンクもありません。そのため、公式に確認できるのは「用途・機能・上限値」であり、「前世代比 X% 向上」といった性能主張ではありません。 25

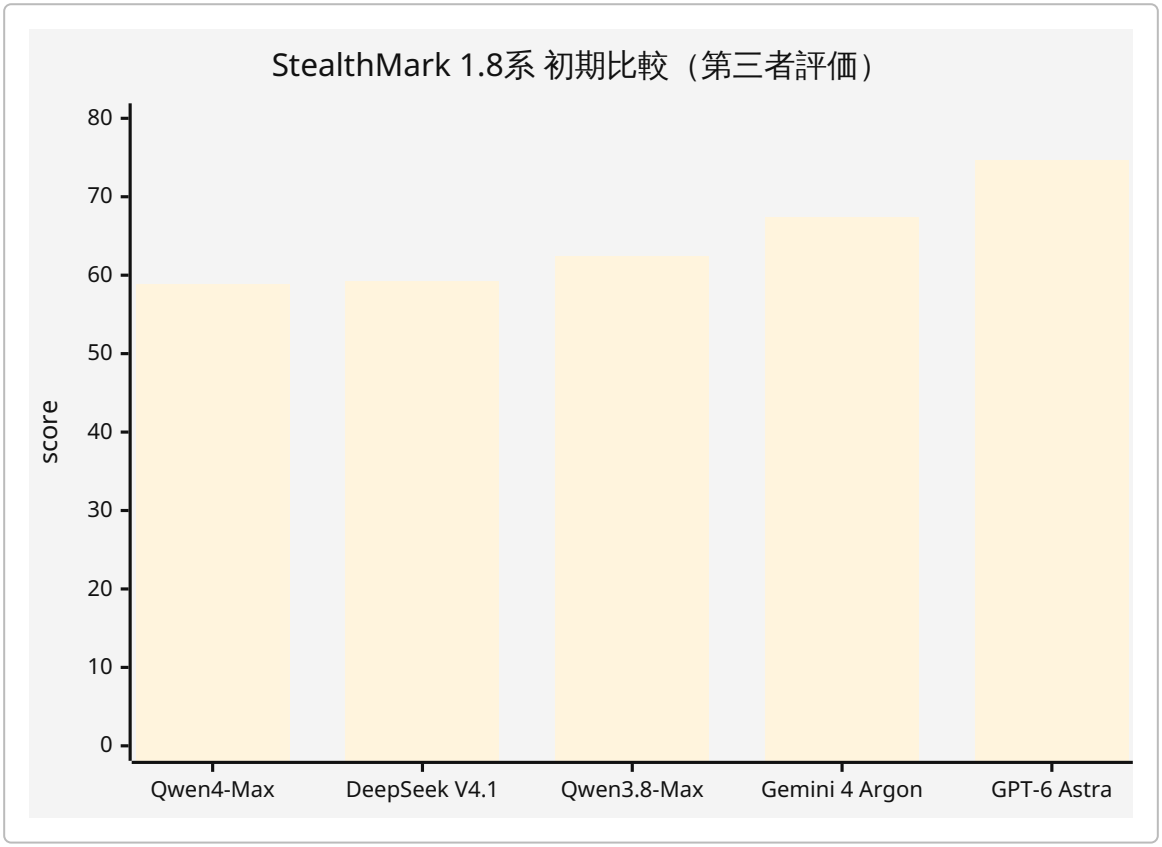
これは Qwen3.8-Max や DeepSeek V4.1 と対照的です。DeepSeek は Terminal-Bench、DeepSWE、HLE、Automation-Bench などの個別スコアを公開しており、Qwen3.8 も model card / GitHub 上に benchmark 情報を持ちます。 26

第三者の同一評価系で比較

Stealth Models の StealthMark 1.8 系では次の初期値が確認できます。これは独自ベンチマークであり、業界標準 benchmark の代替ではありません。 27

Model	StealthMark	General	Visual	情報源信頼度
GPT-6 Astra	74.6	73.4	76.9	第三者・中
Gemini 4 Argon	67.4	63.6	74.4	第三者・中
Qwen3.8-Max-0902	62.4前後	—	—	第三者・中
DeepSeek-V4.1-Flash-1017	59.2	56.8	63.5	第三者・中
Qwen4-Max-Preview	**58.8**	**54.6**	**66.3**	第三者・初期値

[28]



このデータから「Qwen4 が Qwen3.8 より弱い」と一般化するのも早計です。StealthMark は物理・空間 reasoning、長期 planning、SVG/visual construction など独自課題の集合であり、coding repository work、長文 office task、web-agent 等を包括的に測るものではありません。加えて Qwen4-Max-Preview は公開後数日の Preview build です。 [29]

主要競合との製品仕様比較

Model	Context	Max output	Vision	Tool use	価格形態
Qwen4-Max-Preview	1M	393,216	image+video	FC/web/tools	Token Plan Creditsのみ

Model	Context	Max output	Vision	Tool use	価格形態
Qwen3.8-Max	1M	128K前後	image+video	○	$2 in / $6 out per MTok
DeepSeek V4.1 Flash-1017	1M	384K	native multimodal	○	$0.30/$1.20 peak、off-peak半額
GPT-6 Astra	1.05M	128K	image	functions/web/file/computer	$10/$50 per MTok
Gemini 4 Argon	約1M級	最大長decode対応	multimodal	Google grounding等	11月末まで $2/$10、その後 $4/$20

Qwen4-Max、Qwen3.8、DeepSeek、GPT-6 の値は各社一次資料に基づきます。Gemini 4 Argon の現在の promotional pricing は Google 公式 API pricing に掲載されています。[30]

Qwen4-Max の最大の仕様上の差別化点は **393,216-token completion budget** で、GPT-6 Astra の 128K や Qwen3.8-Max の 128K を大きく上回ります。ただし reasoning が同じ予算を使うため、393K の「最終回答本文」が必ず生成できるという意味ではありません。[31]

利用者評価

公開から数日しか経っていないため、利用者評価は非常に初期的です。Hacker News では「No benchmarks yet」という指摘があり、別ユーザーは Qwen Code で xhigh thinking を試し、速度を「okay」と評価しています。これらは個人の体感であり再現性のある測定ではありません。[32]

Reddit では、短い brainstorming prompt からモデルが大量の code / interactive artifact を自主生成するなど、**過剰に積極的な agentic behavior** を報告する利用者がいます。一方、別スレッドでは tool schema の boolean が string 化されるという tool-call 問題が報告されています。いずれも個別環境の anecdote であり、Alibaba が公式に root cause を確認した事実ではありません。[33]

Qwen Code の GitHub issue では、provider-side と思われる tool URL 用中国語 instruction を Qwen4-Max-Preview が prompt injection と誤認し、reasoning や user-facing response で繰り返し言及する、という再現報告があります。これも issue report であって公式確定障害ではありませんが、agent production 評価では確認すべき挙動です。[34]

二次情報の信頼度評価としては、Stealth Models は同一 harness の比較という点で「中」、GitHub issue は再現ログ付きなら「中〜低」、Reddit/HN の感想は「低」、SEO 系レビュー記事は公式仕様の再掲には使えても性能評価の根拠としては「低」とするのが妥当です。

安全性、プライバシー、ライセンス、利用制限

Alibaba Cloud Model Studio の International data retention policy では、通常 inference の prompts / responses は原則 **30日以内**の保持とされます。stateful feature の conversation/files はユーザー削除まで保持される場合があります。[35]

Alibaba Cloud は、別途同意がない限り business data / Member Content を foundation model の training、fine-tuning、improvement に利用しないとしています。Team Token Plan についても conversation data を model training に使用しないことが明記されています。 [36]

International enterprise API customer には ZDR（Zero Data Retention）の仕組みがありますが、**Qwen4-Max-Preview / Token Plan が ZDR eligible model として明示されていることは今回確認できませんでした**。したがって、Qwen4-Max-Preview について「ZDR 対応」と断定するのは避けるべきです。ZDR 対応モデルでは prompt/response を durable storage に書き込まず、処理に必要な encrypted cache のみ一時利用する仕様です。 [37]

Content safety moderation は ZDR と独立しており、完全に無効化する switch は提供されないと Alibaba は説明しています。自動安全フィルターで flag された content は法的・権利保護用途等のため最大2年間保持される場合があります。 [38]

ライセンスについて、Qwen4-Max-Preview は downloadable weights の open-source model としては確認されていないため、Apache-2.0 等の「モデル重みライセンス」を適用できる根拠はありません。利用は Alibaba Cloud Model Studio の service terms に従います。Alibaba は Output の知的財産権を自ら主張しないとし、利用者は適用法・契約・ルールの範囲で Input / Output を利用できます。一方、Model Studio 自体の resale、reverse engineering、Alibaba と競合する product/service の training に Model Studio やその output を使うこと等は禁止されています。 [39]

そのため「商用利用可否」は二分して考える必要があります。**生成物を業務で利用すること自体は契約上可能**ですが、Qwen4-Max-Preview を Token Plan key で自社 SaaS の backend inference engine として組み込むことは現在の Token Plan scope に反します。 [40]

確認済み事実と未確認情報

内容	状態
Qwen4-Max-Preview は公式に存在	**確認済み**
Model ID `qwen4-max-preview`	**確認済み**
1M context	**確認済み**
393,216 max completion budget	**確認済み**
image/video input	**確認済み**
web/tool/function calling	**確認済み**
Preview は11月25日終了	**確認済み**
Token Plan 限定	**確認済み**
International は Singapore	**確認済み**
通常 backend 利用禁止	**確認済み**
正式 Qwen4-Max の発売日	**未確認**
Qwen4-Max の parameter count	**未確認**
MoE / dense 等の architecture	**未確認**

内容	状態
Open weights	**未確認／公開確認なし**
GPU/VRAM要件	**未確認**
Alibaba公式 benchmark scores	**未確認／公開確認なし**
Qwen4-Max が Qwen3.8-Max を総合的に上回る	**未確認**
Preview後に同一モデルが stable 化	**未確認**
Qwen4-Max Preview固有のZDR対応	**未確認**

用途別の採用判断と検証プロンプト

業務内 coding / agent 作業なら、Qwen4-Max-Preview は積極的に評価してよい候補です。特に巨大 repo、長い specification、複数 image/video をまたぐ分析、長時間 agent session では 1M context と大きい output budget の価値があります。まず `reasoning_effort=medium` を標準にし、難しい debugging / architecture / formal reasoning だけ `xhigh` に上げるのが Credits と応答時間のバランスを取りやすい設定です。`low` は分類、軽い変換、単純 code edit などに向きます。effort の値自体は公式仕様ですが、この使い分けは本レポートの運用上の推奨です。 [14]

研究用途では非常に興味深い対象です。特に 1M context retrieval、long-horizon planning、visual-agent loop、reasoning-effort scaling の比較実験に向きます。ただし Preview が短命なので、論文や再現実験ではモデル ID、実行日時、effort、prompt、tool harness、region を必ず固定記録すべきです。Preview が11月25日に retire されるため、後日再実行できない可能性があります。 [14]

プロダクト組込みは現時点では推奨できません。能力以前に Token Plan 利用規約が custom backend を禁止しており、モデル自体も11月25日に retire 予定です。本番なら Qwen3.8-Max の通常 API、DeepSeek API、Gemini API、OpenAI API 等と比較する方が適切です。 [41]

機密データでは、Team plan と privacy terms を優先し、保持ポリシー・地域・ZDR eligibility を契約前に Alibaba に書面確認するのが安全です。Qwen4-Max-Preview 固有の ZDR 対応は今回確認できていません。 [42]

検証では、単一の「賢さ」スコアより、自分の workload を20〜50ケース程度固定して Qwen3.8-Max や競合と blind comparison する方法が有効です。以下は Qwen4-Max-Preview の性質を確認しやすい具体例です。

長文・needle retrieval

「添付した複数文書から、`Project Helios` に関する契約上の義務、例外条項、期限、相互矛盾をすべて抽出してください。各結論について文書名と該当箇所を示し、文書間で矛盾する場合は勝手に解決せず `CONFLICT` としてください。」

repository-level coding

「この repository 全体を調査し、authentication flow を図式化してください。その後、session fixation が成立する経路をコード上の根拠とともに特定し、既存 API compatibility を壊さない最小 patch と regression tests を提案してください。変更前に必ず dependency graph と影響ファイル一覧を提示してください。」

tool-call robustness

「次の JSON Schema に厳密に従って tool call を行ってください。boolean は JSON boolean、array は JSON array として出力し、文字列化しないでください。tool 実行後に schema violation がないか自己検証してください。」

この prompt は、公開された初期利用者報告にある boolean/stringification 系の問題を検証するためのものです。報告自体は未確認の anecdote です。 [43]

agent autonomy / over-action

「まず計画だけを作成し、ファイル変更・コード生成・tool call は一切実行しないでください。計画を10項目以内で提示し、私が `EXECUTE` と返すまで待機してください。」

これは短い相談から大量 artifact を自発生成するという利用者報告が自環境でも起こるかを見る試験です。 [44]

1M-context位置耐性

「この長文 corpus に5つの固有なキーを異なる位置に埋め込みました。キーをすべて抽出し、それぞれの前後500文字を要約し、キー間の因果関係がある場合のみ説明してください。見つからないキーを推測して補完してはいけません。」

vision + reasoning

「添付した20枚の dashboard screenshot を時系列に並べ、KPI の異常点を特定してください。数値をOCR推測で補わず、読めない値は `unreadable` としてください。その上で異常の候補原因を、観測事実と仮説に分離して JSON で返してください。」

reasoning-effort A/B

同じ prompt を `low / medium / xhigh` で各5回実行し、正解率、tool-call 成功率、completion tokens、Credits、wall-clock time を記録します。Qwen4-Max-Preview は effort level を公式にサポートしているため、実際の workload で品質とコストの曲線を作るのが最も有益です。 [14]

主要一次情報と確認日

以下はいずれも **確認日 2026年11月15日（Asia/Tokyo）** です。

ソース	種別	URL
Qwen4-Max-Preview model page	Alibaba Cloud 公式	https://www.alibabacloud.com/help/en/model-studio/qwen4-max-preview
Qwen4-Max-Preview / China docs	Alibaba Cloud 公式	https://help.aliyun.com/en/model-studio/qwen4-max-preview

ソース	種別	URL
Model releases	Alibaba Cloud 公式	https://docs.modelstudio.console.alibabacloud.com/en/model-studio/newly-released-models
Token Plan Personal	Alibaba Cloud 公式	https://www.alibabacloud.com/help/en/model-studio/token-plan-personal-overview
Token Plan Team	Alibaba Cloud 公式	https://docs.modelstudio.console.alibabacloud.com/en/model-studio/token-plan-team-overview
Token Plan usage policy	Alibaba Cloud 公式	https://help.aliyun.com/en/model-studio/token-plan-permissions-and-usage
Token Plan quick start	Alibaba Cloud 公式	https://www.alibabacloud.com/help/en/model-studio/token-plan-team-quickstart
Base URL / endpoints	Alibaba Cloud 公式	https://help.aliyun.com/en/model-studio/base-url
Data retention policy	Alibaba Cloud 公式	https://www.alibabacloud.com/help/en/model-studio/data-retention-policy
Product Terms	Alibaba Cloud Legal	https://www.alibabacloud.com/help/en/legal/latest/alibaba-cloud-international-website-product-terms-of-service-v-3-8-0
Qwen3.8 official repository	Qwen 公式 GitHub	https://github.com/QwenLM/Qwen3.8
Qwen3.8-Max repository	Alibaba Cloud 公式 GitHub	https://github.com/AlibabaCloud-Official/Qwen3.8-max
OpenAI GPT-6 Astra model page	OpenAI 公式	https://developers.openai.com/api/docs/models/gpt-6-astra
Gemini API pricing	Google 公式	https://ai.google.dev/gemini-api/docs/pricing
DeepSeek pricing	DeepSeek 公式	https://api-docs.deepseek.com/quick_start/pricing/
DeepSeek release log	DeepSeek 公式	https://api-docs.deepseek.com/updates/
StealthMark Qwen4 result	第三者 benchmark	https://stealthmodels.com/benchmarks/models/qwen4-max-preview
StealthMark methodology	第三者 benchmark	https://stealthmodels.com/stealthmark
Hacker News discussion	利用者投稿	https://news.ycombinator.com/item?id=50451192

総括すると、**Qwen4-Max-Preview は「存在するが、まだ製品化された Qwen4-Max ではない」**という理解が最も安全です。公式に確定している強みは 1M context、393K completion budget、native vision/video、deep reasoning、豊富な tools/API compatibility です。一方で、architecture、parameter count、公式 benchmark、本番価格、正式版 roadmap は未公表です。そして現時点で採用判断を左右する最大要因は、

性能より **2026年11月25日で retire する期間限定 Preview と、Token Plan の interactive-only 利用制限**です。 45

---

1 3 5 6 14 25 30 31 45 qwen4-max-preview (Latest Model) - Alibaba Cloud Model Studio - Alibaba Cloud Documentation Center
https://www.alibabacloud.com/help/en/model-studio/qwen4-max-preview?utm_source=chatgpt.com

2 16 24 41 Token Plan permissions and usage policies-Alibaba Cloud Model Studio(Model Studio)-阿里云帮助中心
https://help.aliyun.com/en/model-studio/token-plan-permissions-and-usage?utm_source=chatgpt.com

4 23 28 Qwen4-Max-Preview Benchmarks and Game Replays
https://stealthmodels.com/benchmarks/models/qwen4-max-preview?utm_source=chatgpt.com

7 9 Qwen3.8-max/README.md at main · AlibabaCloud-Official/Qwen3.8-max · GitHub
https://github.com/AlibabaCloud-Official/Qwen3.8-max/blob/main/README.md?utm_source=chatgpt.com

8 qwen3.8-max Model Info-Alibaba Cloud Model Studio(Model Studio)-阿里云帮助中心
https://help.aliyun.com/en/model-studio/qwen3-8-max?utm_source=chatgpt.com

10 GitHub - QwenLM/Qwen3.8: Qwen3.8 is the large language model series developed by Qwen team, Alibaba Group. · GitHub
https://github.com/QwenLM/Qwen3.8?utm_source=chatgpt.com

11 17 18 Token Plan (Personal Edition) - Alibaba Cloud Model Studio - Alibaba Cloud Documentation Center
https://www.alibabacloud.com/help/en/model-studio/token-plan-personal-overview?utm_source=chatgpt.com

12 Token Plan (Personal Edition)-Alibaba Cloud Model Studio(Model Studio)-阿里云帮助中心
https://help.aliyun.com/en/model-studio/token-plan-personal-overview?utm_source=chatgpt.com

13 Quick start - Alibaba Cloud Model Studio - Alibaba Cloud Documentation Center
https://www.alibabacloud.com/help/en/model-studio/token-plan-team-quickstart?utm_source=chatgpt.com

15 Token Plan overview - Alibaba Cloud Model Studio - Alibaba Cloud Documentation Center
https://www.alibabacloud.com/help/en/model-studio/token-plan-overview?utm_source=chatgpt.com

19 Overview - Alibaba Cloud Model Studio
https://docs.modelstudio.console.alibabacloud.com/en/model-studio/token-plan-team-overview?utm_source=chatgpt.com

20 Token Plan product updates-Alibaba Cloud Model Studio(Model Studio)-阿里云帮助中心
https://help.aliyun.com/en/model-studio/token-plan-product-updates?utm_source=chatgpt.com

21 Token Plan usage and quota-Alibaba Cloud Model Studio(Model Studio)-阿里云帮助中心
https://help.aliyun.com/en/model-studio/token-plan-usage-quota?utm_source=chatgpt.com

22 API compatibility upgrades - Alibaba Cloud Model Studio
https://docs.modelstudio.console.alibabacloud.com/en/model-studio/api-compatibility-upgrade-notice?utm_source=chatgpt.com

26 Change Log | DeepSeek API Docs
https://api-docs.deepseek.com/updates/?utm_source=chatgpt.com

27 29 StealthMark: The AI Benchmark Explained | Stealth Models
https://stealthmodels.com/stealthmark?utm_source=chatgpt.com

32 Qwen 4 Max Preview | Hacker News
https://news.ycombinator.com/item?id=50451192&utm_source=chatgpt.com

33 44 Qwen 4 Max preview submarine test
https://www.reddit.com/r/Qwen_AI/comments/1xw18x4/qwen_4_max_preview_submarine_test/?utm_source=chatgpt.com

34 qwen4-max-preview: server-side Chinese tool-URL rule (routify) leaks into context, model flags it as prompt injection and comments on it · Issue #17086 · QwenLM/qwen-code · GitHub
https://proxy.audy.io/v/s-Z2l0aHViLmNvbQ/QwenLM/qwen-code/issues/17086?utm_source=chatgpt.com

35 37 42 Alibaba Cloud Model Studio Data Retention Policy - Alibaba Cloud Model Studio - Alibaba Cloud Documentation Center
https://www.alibabacloud.com/help/en/model-studio/data-retention-policy?utm_source=chatgpt.com

36 Alibaba Cloud Model Studio Data Retention Policy - Alibaba Cloud Model Studio - Alibaba Cloud Documentation Center
https://www.alibabacloud.com/help/en/model-studio/data-retention-policy.md?utm_source=chatgpt.com

38 Alibaba Cloud Model Studio Data Retention Policy - Alibaba Cloud Model Studio - Alibaba Cloud ドキュメントセンター
https://www.alibabacloud.com/help/ja/model-studio/data-retention-policy?utm_source=chatgpt.com

39 40 Alibaba Cloud International Product Terms of Service Overview - Legal - Alibaba Cloud - Legal - Alibaba Cloud Documentation Center
https://www.alibabacloud.com/help/en/legal/latest/alibaba-cloud-international-website-product-terms-of-service-v-3-8-0?utm_source=chatgpt.com

43 Qwen4-max-preview has a tool calling problem
https://www.reddit.com/r/Qwen_AI/comments/1xul1hu/qwen4maxpreview_has_a_tool_calling_problem/?utm_source=chatgpt.com